Architektura IBM POWER, do nitra nejvýkonnějšího superpočítače na světě

19. prosince 2018
CIIRC Praha

Milan Král, IBM
Radek Špimr

Rank	System	Cores	Rmax (TFlop/s)	Rpeak (TFlop/s)	Power (kW)
1	**Summit** - IBM Power System AC922, IBM POWER9 22C 3.07GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband , IBM DOE/SC/Oak Ridge National Laboratory United States	2,397,824	143,500.0	200,794.9	9,783
2	**Sierra** - IBM Power System S922LC, IBM POWER9 22C 3.1GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband , IBM / NVIDIA / Mellanox DOE/NNSA/LLNL United States	1,572,480	94,640.0	125,712.0	7,438

CORAL

CORAL

TOP500

The List.

Rank	System	Cores	Rmax (TFlop/s)	Rpeak (TFlop/s)	Power (kW)
1	**Summit** - IBM Power System AC922, IBM POWER9 22C 3.07GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband , IBM DOE/SC/Oak Ridge National Laboratory United States	2,282,544	122,300.0	187,659.3	8,806
2	**Sunway TaihuLight** - Sunway MPP, Sunway SW26010 260C 1.45GHz, Sunway , NRCPC National Supercomputing Center in Wuxi China	10,649,600	93,014.6	125,435.9	15,371
3	**Sierra** - IBM Power System S922LC, IBM POWER9 22C 3.1GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband , IBM DOE/NNSA/LLNL United States	1,572,480	71,610.0	119,193.6	
4	**Tianhe-2A** - TH-IVB-FEP Cluster, Intel Xeon E5-2692v2 12C 2.2GHz, TH Express-2, Matrix-2000 , NUDT National Super Computer Center in Guangzhou China	4,981,760	61,444.5	100,678.7	18,482
5	**AI Bridging Cloud Infrastructure (ABCI)** - PRIMERGY CX2550 M4, Xeon Gold 6148 20C 2.4GHz, NVIDIA Tesla V100 SXM2, Infiniband EDR , Fujitsu National Institute of Advanced Industrial Science and Technology (AIST) Japan	391,680	19,880.0	32,576.6	1,649

CORAL Installation at ORNL

OAK RIDGE National Laboratory

CORAL Installation at LLNL

CORAL

Order of Magnitude Leap in Computational Power

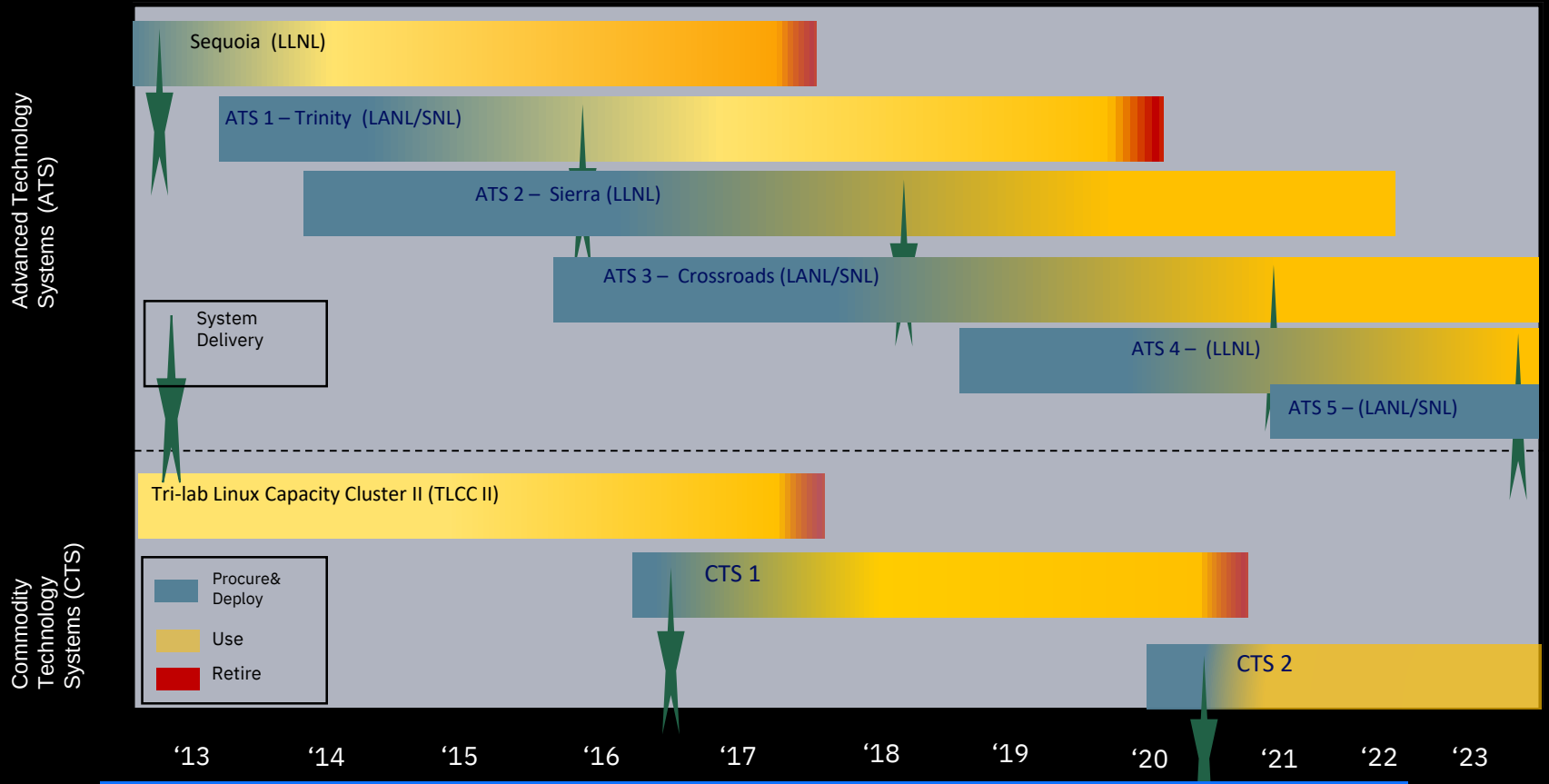
Real, Accelerated Science

3+EFLOPS	10X	5-10X
Tensor Ops	Perf Over Titan	Application Perf Over Titan

- Deployments beginning with full acceptance in 2018
- Significant application performance over Titan (AMD/NVIDIA)
  - Achieved with ¼ the servers

IBM

Sierra is the next ASC ATS platform



Sequoia and Sierra are the current and next-generation Advanced Technology Systems at LLNL

The Sierra system

Compute Node
AC922

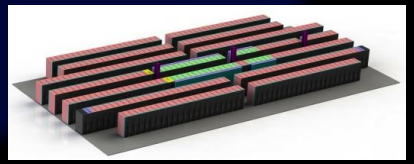
2 IBM POWER9 CPUs
4 NVIDIA Volta GPUs
NVMe-compatible PCIe 1.6 TB SSD
256 GiB DDR4
16 GiB Globally addressable HBM2 associated with each GPU
Coherent Shared Memory

Compute Rack

Standard 19”
Warm water cooling

Compute System

4320 nodes
1.29 PB Memory
240 Compute Racks
125 PFLOPS
~12 MW



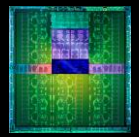
Components

IBM POWER9

- Gen2 NVLink

NVIDIA Volta

- 7 TFlop/s
- HBM2
- Gen2 NVLink



Mellanox Interconnect

Single Plane EDR InfiniBand
2 to 1 Tapered Fat Tree

GPFS File System

154 PB usable storage
1.54 TB/s R/W bandwidth

IBM POWER SYSTEMS

AC922

Designed for the AI Era

Delivering Enterprise-Class AI

High level System Overview

2-Socket, 2U Packaging
40 P9 Processor cores
4/6 NVIDIA Volta 2.0 GPUs
1/2 TB Memory (16x - 64GB DIMMs)
4 PCIe Gen4 Slots
2x SFF (HDD/SSD), SATA, Up to 7.7 TB storage
Supports 1.6TB and 3.2TB NVMe Adapters
Redundant Hot Swap Power Supplies and Fans
Default 3 year 9x5 warranty, 100% CRU

POWER9

Slim core

20 CPU core, SMT=4

Linux, HPC, & KVM

Fused core

12 CPU core, SMT=8

PowerVM

IBM LC

Models

OpenPOWER

AC922

2018

IBM Scale-Out

AIX + IBM i + Linux

2018

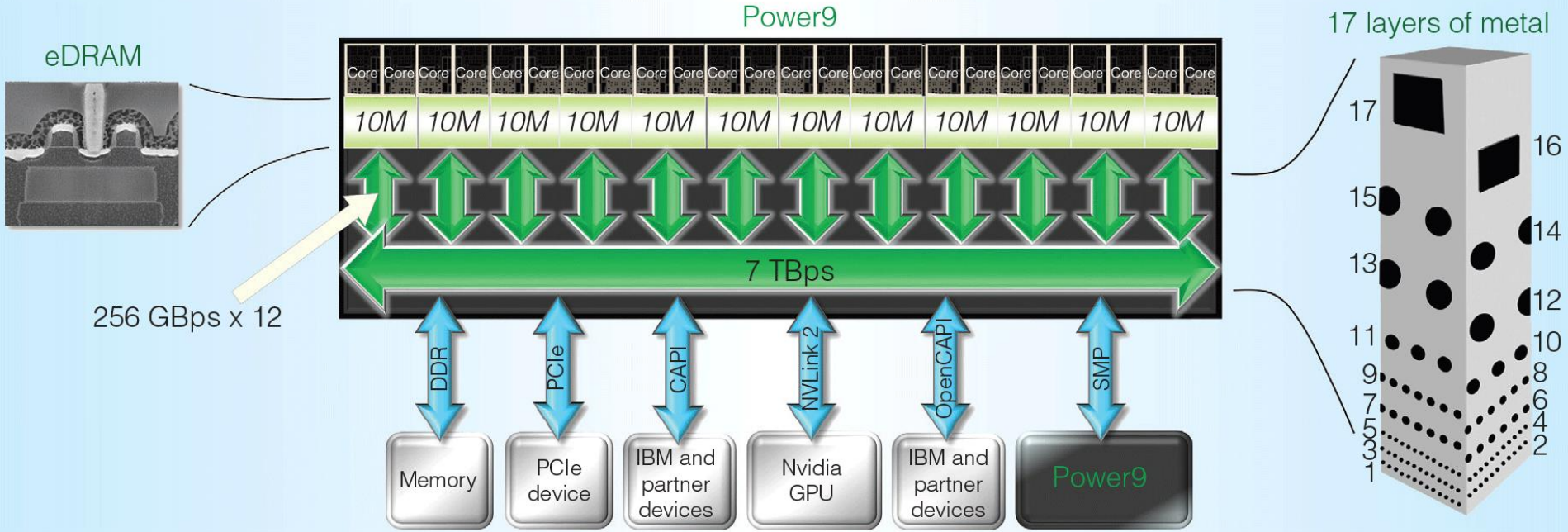
IBM Enterprise

Centaur RAM

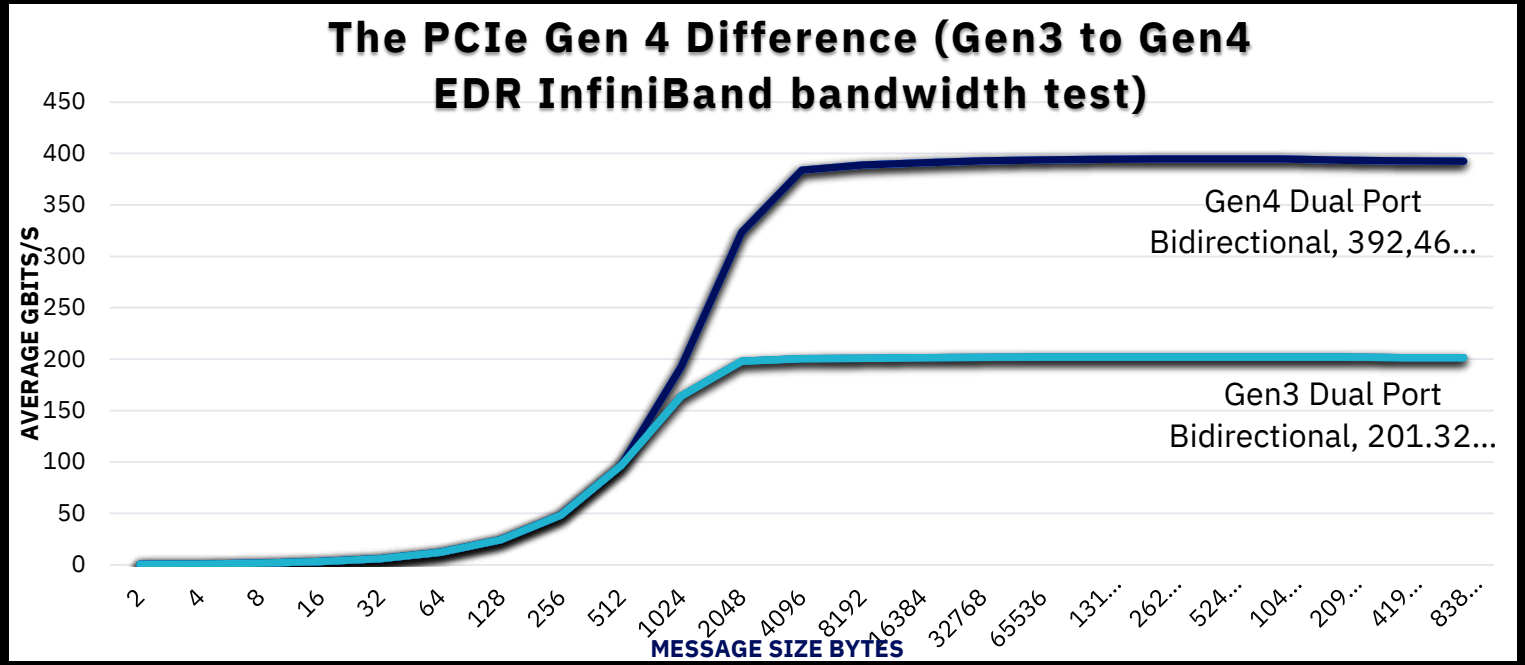
AIX + IBM i + Linux

IBM

POWER9: Konektivita



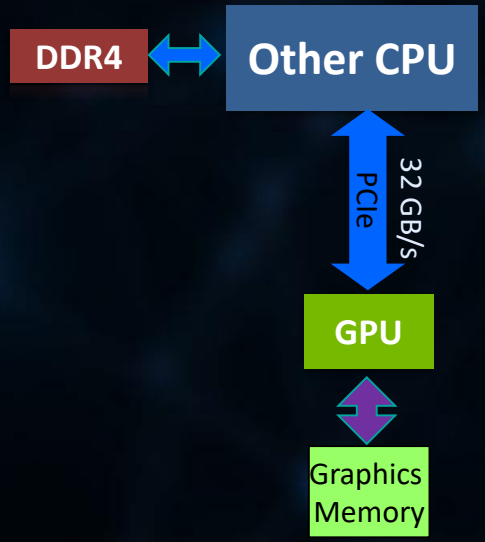
InfiniBand EDR 100Gb/s – PCIe Gen 4 verses PCIe Gen 3



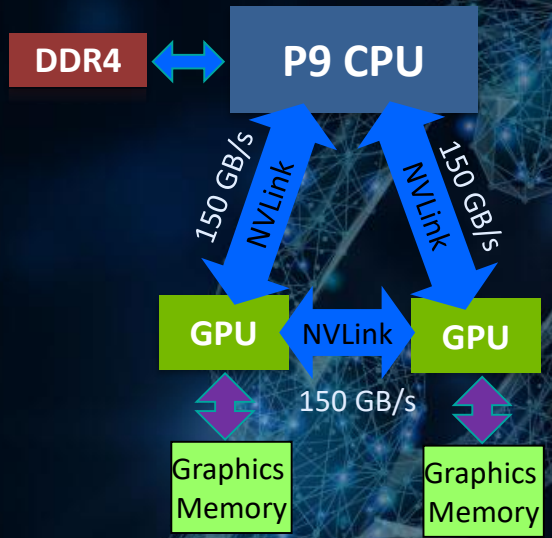
- ~2x faster IB network connectivity enabled PCIe Gen 4
- Best server technology for clustering servers with PCIe Gen 4 networking and will be expanding to other devices to leveraged the aggregate bandwidth advantages

Architektura CPU s GPU

CPU with PCIe link to GPU



POWER9 with NVLink 2.0 Volta Technology



CORAL Software

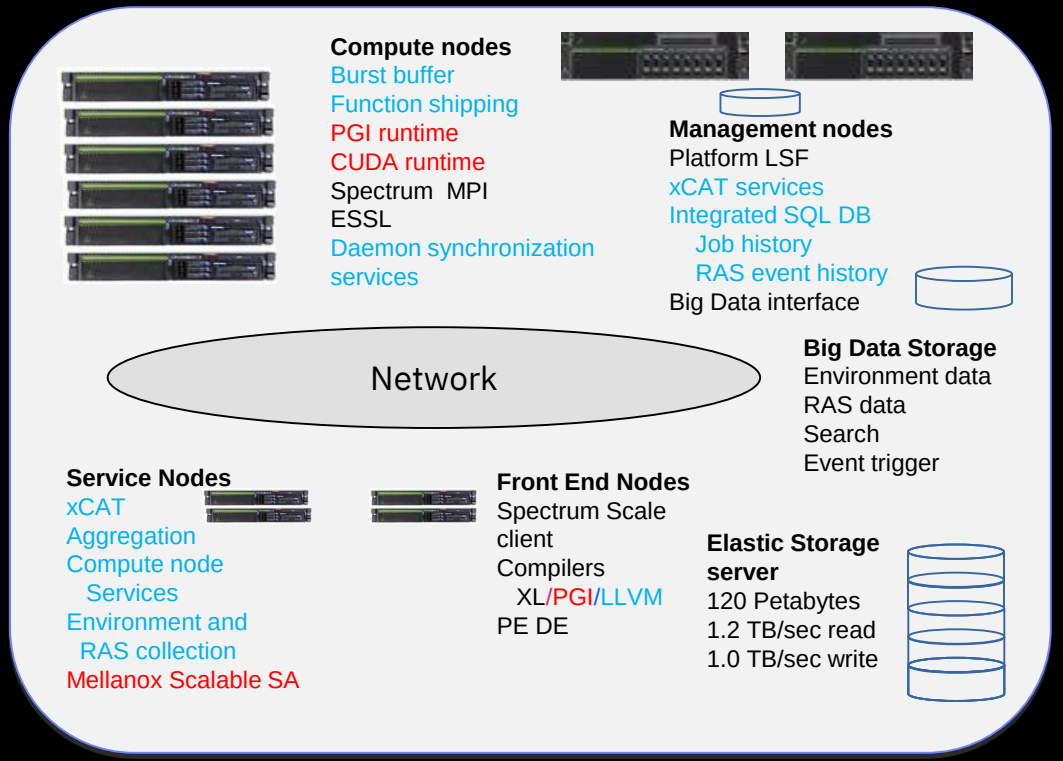
CORAL Software Stack leverages IBM's Mainstream Software Roadmap

Additional Innovations include

- GPU Accelerated Math Libraries
- Spectrum Scale Scaling
- Messaging layer to exploit GPU Direct
- Burst Buffer and Function Shipping
- Scalable Cluster System Management Infrastructure and integrated with LSF
- Open Source Tools, Scalable Diagnostics
- Accelerated Support for a Diverse Set High Level Programming Models

CORAL Software Architecture

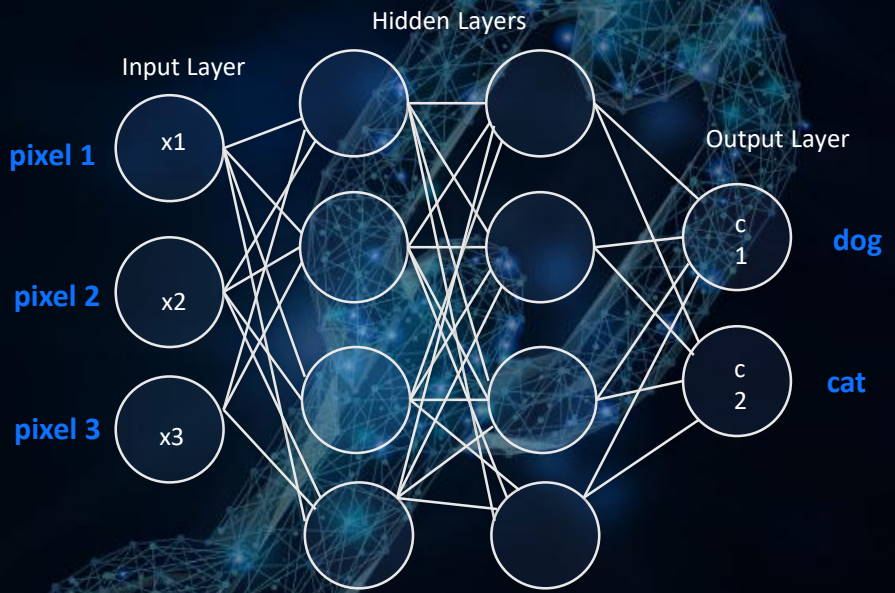
CORAL Innovations



Summit Applications

- **QMCPACK:** Material Science,
- Analyzing materials on atomic level
- **CoMet:** Genome Analysis
- **MENNDL:** AI Deep Learnin Networks
- **GTC:** Gyrokinetic Toroidal Code: Plasma Simulaton
- **NAMD:** Computational Biophysics
- Molecular Dynamic Simulation,
- **NUCCOR:** Nuclear Physics
- **RAPTOR:** Engineering/Combustion CFD
- **GronOR:** Comp. Chemistry, Photovoltaic Apps.

POWER9 AC922 a AI/ML/DL



- **POWER9 delivers 3.8x reduction in AI training with same NVIDIA GPU**

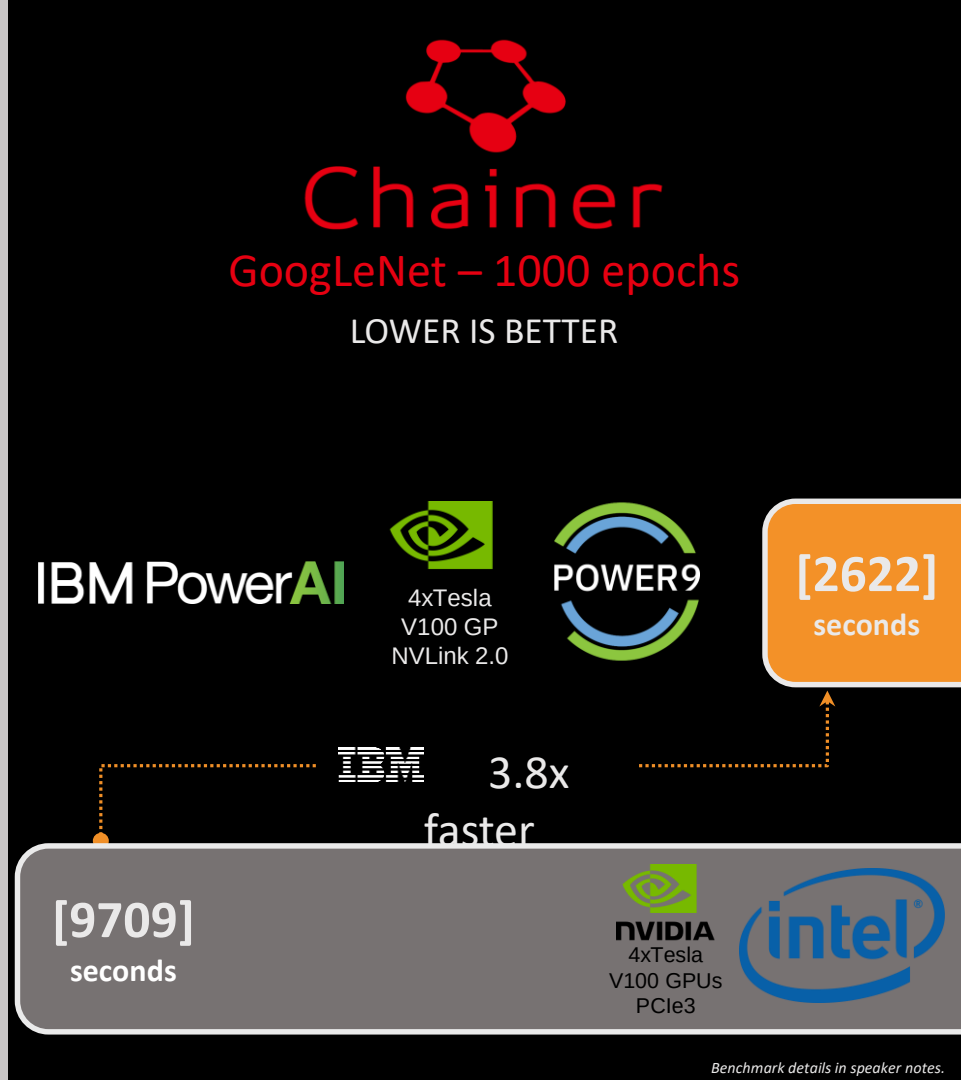
train more | build more | know more

Critical capabilities (regression, nearest neighbor, recommendation systems, +++) operate on more than just the GPU memory

Use Server and GPU memory to support higher resolution data by moving large amounts of data between the CPU and GPU

PowerAI automatically enables seamless use of Server and GPU memory

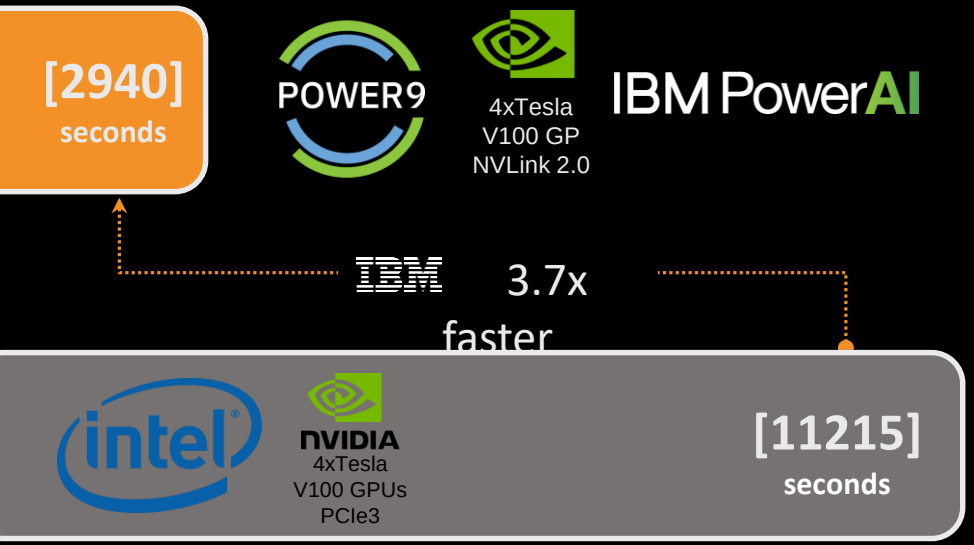
NVLink 2.0 and POWER9 significantly cuts training times and boosts performance (accuracy) of the model with higher resolution data



Caffe

GoogLeNet – 1000 epochs

LOWER IS BETTER



Benchmark details in speaker notes.

POWER9 delivers 3.7x reduction in AI training with same NVIDIA GPU

train more | build more | know more

Critical capabilities (regression, nearest neighbor, recommendation systems, +++) operate on more than just the GPU memory

Use Server and GPU memory to support higher resolution data by moving large amounts of data between the CPU and GPU

PowerAI automatically enables seamless use of Server and GPU memory

NVLink 2.0 and POWER9 significantly cuts training times and boosts performance (accuracy) of the model with higher resolution data

- **POWER9 delivers 2.3x more images processed per second vs tested x86 systems**

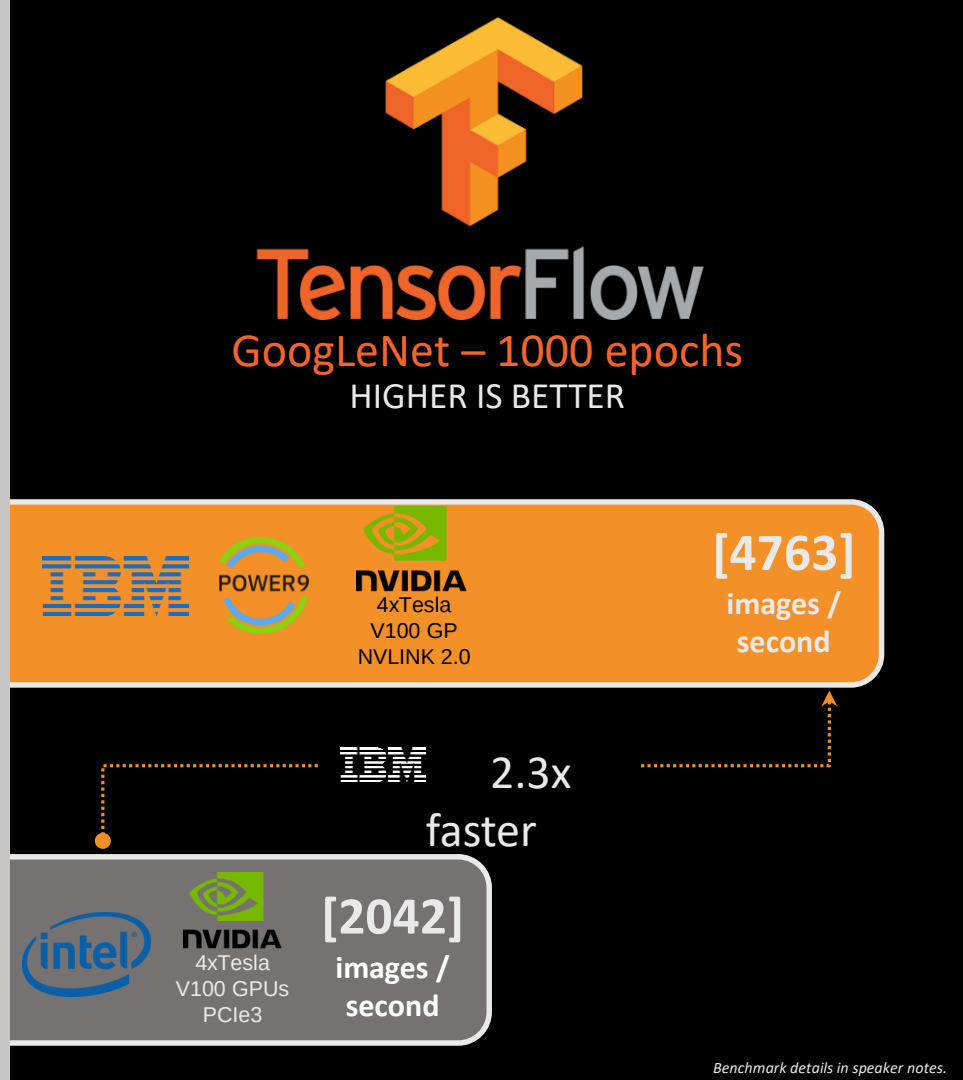
train more | build more | know more

Critical capabilities (regression, nearest neighbor, recommendation systems, +++) operate on more than just the GPU memory

Use Server and GPU memory to support higher resolution data by moving large amounts of data between the CPU and GPU

PowerAI automatically enables seamless use of Server and GPU memory

NVLINK 2.0 and POWER9 significantly cuts training times and boosts performance (accuracy) of the model with higher resolution data




Děkujeme za pozornost